Defending Against Neural Fake News

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, Yejin Choi

University of Washington, AI2

@rown

Motivation

Online disinformation, or fake news intended to deceive, has emerged as a major societal problem. Currently, fake news articles are written by humans, but recently-introduced AI technology might enable adversaries to generate fake news. **Our goal is to reliably detect this "neural fake news" so that its harm can be minimized.**

We take inspiration from the security community. Modern computer security relies on careful ***threat modeling***: identifying potential threats and vulnerabilities from an adversary's point of view, and exploring potential mitigations to these threats.

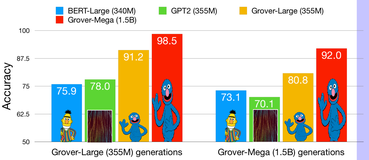
Grover: a threat model for Neural Fake News

A news article is composed of many fields. Grover is a new model for controllable generation that allows for one field to be generated conditioned on any set of context fields.

Switching the headline lets an adversary easily generate neural fake news.

https://grover.allenai.org

Defending against Grover



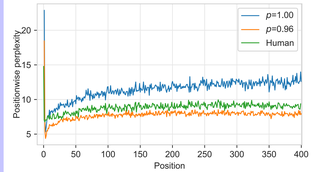
Left-to-right generators are the best at telling apart Grover generated text as real or fake. Even controlling for domain adaptation, they outperform bidirectional models.

Controllably generating fake news is possible. But, the generated articles are easily classified as fake by other generators.

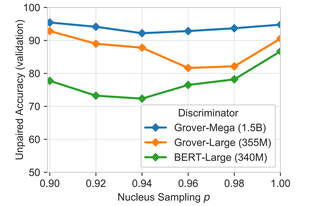
Paper, data, and more at rowanzellers.com/grover

Why is Grover-written text easy to spot?

Exposure bias manifests with length:



Yet top-p (or top-k) sampling also leaves artifacts visible to models that have similar structure to the generator.



There is a sweet spot of careful variance reduction, yet, discrimination is still easy.

More work must be done on generalization to unseen adversaries. However, Grover can classify GPT-2 mega generated news as fake with 96% accuracy. Grover is also robust to simple attacks such as rejection sampling, if it can train on in-domain samples.

Safe release

In our paper, we introduced Grover, a state-of-the-art model for detecting neural fake news. However, because of the underlying mechanics of current text generation systems, strong disinformation detectors will also be strong disinformation generators.

We publicly released Grover-Large after arXiving the paper, and released Grover-Mega to researchers who signed a release form. Currently, Grover-Mega is freely available.

However, Grover is not a panacea. Though in our experiments we found Grover tends to be a highly accurate discriminator of neural fake news, its performance might degrade in practice; moreover, there are serious consequences to both false negatives and false positives.

Our research is the first step toward studying algorithmic defense mechanisms against mass production of fake news by machines. We invite follow up research on this topic, which we also intend to do.